R E P O R T R E S U M E S

ED 019 904 FL 000 824

A CLASSIFICATION AND RETRIEVAL SYSTEM FOR RECORDED FOREIGN LANGUAGE TAPES.

BY- KAY, J.B. JAMESON, A.

PUB DATE NOV 67

EDRS PRICE MF-$0.25 HC-$0.44 9P.

DESCRIPTORS- *CLASSIFICATION, *INFORMATION RETRIEVAL, *TAPE RECORDINGS, *MODERN LANGUAGES, *INSTRUCTIONAL MATERIALS CENTERS, LIBRARY COLLECTIONS, INSTRUCTIONAL MATERIALS, RESOURCE CENTERS, INDEXING, CODIFICATION, CATALOGING, UNIVERSITIES, UNIVERSITY OF ESSEX,

A CLASSIFICATION SYSTEM FOR FOREIGN LANGUAGE RECORDINGS, DEVISED FOR USE AT THE UNIVERSITY OF ESSEX, IS DESCRIBED IN THIS ARTICLE. CODING AND CROSS-REFERENCE SYSTEMS FOR THE VARIOUS CATEGORIES OF TAPES ARE OUTLINED, AND SPECIFIC EXAMPLES ARE INCLUDED. APPENDED IS A LIST OF TAPE REFERENCE CODES. THE SYSTEM OUTLINED IS NOT CONSIDERED IDEAL, BUT IT IS OPEN-ENDED AND MEETS THE UNIVERSITY'S ACCESS AND STORAGE NEEDS, AND IS OFFERED AS A POSSIBLE AID TO OTHER INSTITUTIONS FACED WITH SIMILAR NEEDS. THIS ARTICLE WAS PUBLISHED IN "INTERNATIONAL REVIEW OF APPLIED LINGUISTICS IN LANGUAGE TEACHING," VOLUME 5, NUMBER 4, NOVEMBER 1967, PAGES 193-200. (AR)

ED019904

IRAL

International Review of Applied Linguistics in Language Teaching

V/4 (1967)
(November)

JULIUS GROOS VERLAG HEIDELBERG P.O.B. 629

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Julius Groos Verlag

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

© 1963 Julius Groos Verlag

6900 Heidelberg/Germany, Gaisbergstraße 6–8
P. O. Box 629. Cable Address: Groos Heidelberg
Banking Account No. 02/52338 Deutsche Bank Heidelberg
Printed by Julius Beltz, Weinheim/Germany.

Editors: Prof. Dr. Bertil Malmberg, University of Lund/Sweden
Prof. Dr. Gerhard Nickel, Universität Kiel/Germany

Published quarterly
Erscheint vierteljährlich

Subscription per Vol.	Europe (excl. France)	Overseas (excl. USA)
Individual	DM 38.– (US$ 9.50)	DM 39.– (US$ 9.75)
Institutional	DM 42.– (US$ 10.50)	DM 43.– (US$ 10.75)
Single Copies	DM 10.50 (US$ 2.65)	DM 11.– (US$ 2.75)

Postage included
France: Price: NF 60 – Sole Agency: Librairie Marcel Didier – 4 et 6, rue de la Sorbonne – Paris V
USA: Price: US$ 10.75 – Sole Agency: Chilton Books, A Division of Chilton Company – East Washington Square – 525 Locust Street – Philadelphia 6, Penna. (19106)

FL 000 824

A CLASSIFICATION AND RETRIEVAL SYSTEM FOR RECORDED FOREIGN LANGUAGE TAPES

J.B. Kay – A. Jameson

Le *Language Centre* de l'Université d'Essex proposa en 1964 un plan pour les collections de spécimens de langues étrangères enregistrés sur bande magnétique. En 1965 un *Foreign Language Recordings Project* fut nommé, confié à un comité et aux frais de la *Nuffield Foundation*. Des matériaux existent actuellement pour les langues française, allemande, portugaise, russe et espagnole.

Le système de classement utilisé à Essex peut offrir un intérêt pour les autres institutions qui connaissent les mêmes problèmes. Le corpus est d'abord divisé d'après la langue, ayant chacune un chiffre, commençant par 001 selon l'ordre d'accession. La première lettre réfère à la division en langues (F pour le français, etc.). La deuxième est celle entre matériaux littéraires et non-littéraires, la littérature étant à son tour divisée selon les genres: D = théâtre, P = prose, etc. Un enregistrement ayant la cote RV 001 est donc identifié immédiatement comme poésie russe, SD 007 serait compris comme pièce de théâtre espagnol, etc. Pour les enregistrements non-littéraires il a fallu jusqu'ici se contenter de divisions relativement sommaires: chimie, économie, voyages, guerre, etc., qui permettront des subdivisions ultérieures rendues nécessaires par l'augmentation des matériaux. Voir le schéma, p. 00. L'Appendice p. 00 donne un aperçu du système. Le système proposé n'est pas regardé comme un idéal, seulement comme une tentative de maîtriser la situation actuelle à Essex.

Das *Language Centre* der Universität Essex machte im Jahr 1964 einen Vorschlag, auf Tonband aufgenommene Fremdsprachenproben zu sammeln. 1965 wurde ein *Foreign Language Recordings Project* eingerichtet und die Arbeit, auf Kosten der *Nuffield Foundation*, einem Komitee überlassen. Inzwischen existieren Materialien für folgende Sprachen: Französisch, Deutsch, Portugiesisch, Russisch und Spanisch.

Das in Essex benutzte Klassifikationssystem dürfte auch für andere Institutionen mit denselben Problemen von Interesse sein. Das Corpus ist zunächst nach Sprachen eingeteilt, mit Ziffern für jede Sprache: 001 usw. in der Erwerbsreihenfolge. Der erste Buchstabe bezieht sich auf die Sprache (F für Französisch, usw.), der zweite auf die Einteilung in literarische und nicht-literarische Materialien, wobei die Literatur wiederum nach ihren Arten unterteilt ist: D = Theater, P = Prosa, usw. Eine Aufnahme mit der Nummer RV 001 wird also als russische Poesie identifiziert, SD 007 als ein spanisches Theaterstück, usw. Für die nicht-literarischen Aufnahmen hat man sich bisher mit groben Einteilungen begnügen müssen: Chemie, Ökonomie, Reisen, Krieg, usw., die, mit Zunahme des Materials, spätere, endgültige Unterteilungen erlauben werden. Siehe Schema S. 00. Der Anhang S. 00 gibt eine Übersicht des Systems. Das vorgeschlagene System wird nicht als Ideal angesehen, sondern als ein Versuch die jetzige Situation in Essex zu bewältigen.

FL 000 824

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY Julius Groos Verlag

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Among the post-war developments in language teaching has been the increase in the number and variety of technical aids available to teacher and student alike, and notably among these is the tape recorder, which has enabled models of native speech to be presented much more conveniently than previously. At the same time, fresh thinking on the methods of teaching foreign languages has placed much greater emphasis on the spoken language. These two developments have together brought home the realisation that there is a great lack of suitable practice material, especially of a specialised kind and in the restricted fields which are of interest to a large number of adult learners. Obviously, whatever the ultimate aims of the student, in the initial stages of learning a foreign language there is the need for a similar linguistic content, but as the learner progresses much greater attention must be given to the use he wishes to make of the language he is learning, and courses should be appropriately designed for these needs.

It was as a result of thinking along these lines that a proposal was made in 1964 by the Language Centre of the University of Essex to the National Committee on Research and Development in Modern Languages for a scheme to collect samples of foreign language recordings. Recordings of foreign language material had, of course, begun to be acquired as soon as the University began to appoint its first language teaching staff, but it was not until September 1965 that the Foreign Languages Recordings Project became officially designated and a small staff was appointed to run it. The project is financed by the Nuffield Foundation and has as its brief the collection of a corpus of recorded foreign language material in the disciplines studied at Essex. Material is at present being collected in the following languages: French, German, Portuguese, Russian and Spanish, and is of interest then both for its content (to satisfy the needs of university departments) and for its form (for course material).

It was clear at the outset that some form of classification system was already necessary in order to allow students and staff to find out what was available in recorded form, and to permit some system of storage which provided rapid access to a given tape, and the need became more urgent as our collection began to grow.

The classification system outlined below has been evolved with our particular needs in view but it may be interest to other institutions faced with similar problems. We needed a convenient short coding for identification of tapes and a detailed index in some accessible form of what each tape contained. No originality is claimed for the coding system in use here which is simply a two-letter prefix followed by a three-figure serial number.

The first cut in the corpus of material is by language, and each language has its own series of numbers starting from 001 and deriving from chronological order of accession or processing. The first *letter* of the reference coding is therefore the initial letter of the language we are dealing with: F for French, G for German, etc. Possible later ambiguities may be removed by adding one or more

lower case letters (for example, Danish material would be designated Da and Dutch Du) but at present no such difficulties exist here because of the limited range of languages required by our degree course structure.

The second major cut in the collection is between literary and non-literary material. This somewhat arbitrary division was prompted both by greater accessibility of the former material compared with the second and by the demands of register classification. We have found it convenient to use this division to provide us with two separate card-index systems suitably cross-referenced. (1) An Index by Authors & (2) An Index by Symbols. The Index by Authors forms a literary catalogue in which the material is arranged alphabetically by the author's surname in the conventional manner. In addition, in the opposite top corner of each index card is the tape reference coding which contains the formal category, for literature is sub-divided into formal categories which are indicated by the second letter of the reference coding. Again the notation is as broad and as simple as possible: D for Drama, P for Prose, S for Song and V for Verse. The second letter of the reference for literature is therefore mnemonic like the first letter denoting language. Thus a tape coded RV 001 would be immediately identified as a recording of Russian poetry; similarly SD 007 would be a recording of a Spanish play, FS 027 a collection of French songs.

In the literary index, the classification is strictly formal, so that Eugeni Onegin (described by Pushkin as a "novel in verse") is RV, while dramatised prose works, for which the Russians have special weakness, are RD. On the other hand some items which consist of passages extracted from standard French prose works are labelled 'FP' despite the fact that they have been dramatised for recording by the use of a number of speakers for the characters and of sound effects (e. g. extracts from Hugo's *Les Misérables*).

Under literature we include recordings of criticism in literary form, reminiscences and memoirs, etc., so long as the material is a reading and is not being composed as it is delivered. Some inconsistencies have been unavoidable. For example, the Russian comedian Arkadi Raikin has a number of recordings consisting of sketches and longer pieces in dramatic form which we classified RD and indexed under Raikin, although not written by him. This course has been adopted because no-one would think of looking for a Raikin piece under the name of the relatively obscure writer of his sketch.

For purposes of cross-reference, each author has a card on which are entered all substantial references to him in the work of other people in any other recordings we have, and this card is arbitrarily styled an Author card. Similarly, in the case of anthologies and collections which cannot be conveniently broken up and recorded under particular authors there will be special cards filed under Anthology and giving full details of the contents, with separate entries on the relevant author cards. Subject matter of the items in the literary index is not separately entered, a recognition of the primarily formal nature of literature.

When we came to consider the non-literary material we were faced by much more difficult problems. For consistency, it was again decided to adopt the principle that one card should contain the full details of the tape and that there should be cross-references in abbreviated form on cards in general categories related to subject. However, a way needed to be found of classifying non-literary material such that information about both its subject matter and its form was provided. The index eventually decided on is divided into a number of categories distinguished by letter symbols, together with a number of subject or topic cards. The subject cards themselves cover fairly broadly defined fields which will allow later sub-division as required, and, of course, they may be added to as new incoming material may demand. Among the subject headings so far used are the following: Chemistry, Course Material, Economics, Historic Occasions, History, Life and Work, Linguistics, Medicine, Performing Arts, Philosophy, Physics, Politics, Travel, War and Military. A full list is given in the appendix.

It was in coming to a decision about dividing the material into formal categories that we found ourselves attempting something infrequently done. In the end we have analysed our material by reference to two fairly arbitrarily chosen parameters, namely that of the *number* of participants in a speech event, that is one or more than one, and that of the *degree of preparedness* for the event. From these two axes it is possible to construct a grid covering most types of spoken (non-literary) material.

Classification of spoken non-literary recordings by formal criteria

Monologue	1. Unprepared (novel speech) Z	2. Prepared but unscripted (exposition) Y	3. From notes (lecture) X	4. Fully prepared (address, etc.) W
Polylogue	5. Unprepared (free discussion) T	6.	7. One speaker prepared (interview) R	8. Both or all speakers prepared (debate) Q

In practice, cell 1 is quite rare and the result of a somewhat artificial situation, e.g. the disembodied voice of the radio commentator. It can also be elicited by asking people questions requiring off-the-cuff exposition, such as 'Can you describe how you put on your overcoat?' or 'Can you explain to me how to

make a call from a public telephone box?' There is certain to be a *degree* of preparation even for this type of exercise in so far as the speaker must rapidly marshal his thoughts into some kind of logical order. Cell 2 provides in fact for the general type of monologue, since most people who embark on monologues have a predetermined line of thought to pursue, as, for example, in describing a process (exposition), telling a joke or a story, giving a series of political exhortations, and so on. Cell 3 follows Cell 2 along the preparedness parameter, the degree of predetermination being increased by the preset *length* of the event. To this category are assigned, for example, all lectures. Cell 4 is what Abercrombie[1]) calls 'spoken prose' and would include the majority of sound broadcasting (e.g. news bulletins), formal set pieces e.g. the Queen's speech from the throne or other things in a similar 'frozen' style.

Turning to the polylogues, Cell 5 is the typical polylogue type which Abercrombie (op. cit.) labels conversation, in which "there is opportunity for give and take". It includes discussion and conversation of all degrees of formality. Cell 6 is blank for the reason that a conversation or discussion in which all the participants knew what they were going to say would be pointless, although it is conceivable that certain radio broadcasts or political meetings might fall within this category. However, in this case there would be no true discussion and the speech event would more correctly be classified as a series of monologues of the type in Cell 2. Cell 7 covers interviews, interrogations, discussions after seminars i.e. discussion groups with a group leader to guide their progress, and so on, while Cell 8 would include debates (before being thrown open to the house), formal disputations, court proceedings, etc.

Having established these seven formal categories it was then necessary to designate a letter code to follow the initial language code letter as in the case of the literary material. A mnemonic code, while desirable, proved impracticable; so a series of arbitrary letters was chosen from those not already in use for the literary classes, and these are shown in the grid. Thus a tape coded FR 039 would be immediately known to be an interview or discussion in French in which one of the participants had prepared his questions in advance. Similarly, RX 047 would indicate a tape consisting of a lecture prepared with notes and given in Russian. In practice we add the subject classification in abbreviated form after tape coding, e.g. RX 047 Hist. (history) or SX 024 Ling. (linguistics). The grid thus received the letters: Cell 1 : Z, Cell 2 : Y, Cell 3 : X, Cell 4 : W, Cell 5 : T, Cell 6 : –, Cell 7 : R, Cell 8 : Q.

The classification of non-literary material was largely devised in this formal way to make linguistic analysis easier in the preparation of course materials. It

---

[1]) D. Abercrombie: *Conversation and Spoken Prose,* the first of four public lectures on spoken language given at the University of Ghana in Feb. 1959. Published in *E.L.T.* Vol. XVIII No. 1 1963 and republished in *Studies in Phonetics and Linguistics,* O.U.P. London 1965.

is an open-ended system in so far as there is room for the creation of new categories if these are found to be desirable. For instance, we have added a formal category 'N' which is given to types of material of mixed formal content but which are of frequent occurrence, for example news commentary or documentary which contains extracts from speeches, commentaries, sound effects, interviews and so on. Similarly, the designation 'C' for course material (which is both a 'subject' and a formal category, as its topic is in most cases its own form) is a broad classification which may include pronunciation practice, structure drills, dialogues and so on. Course material may be subdivided as required into various 'topics' which can be shown as a suffix to the tape coding in the same way as the other material in the non-literary index.

A word may not be out of place about the allocation of numbers to the tapes. As a matter of principle, numbering systems are less adaptable to mnemonic use, although we found we soon had to abandon the attempt to find unambiguous letters for the non-literary items. However, the number given to a particular tape is purely an access number and is therefore a check of uniqueness. We have decided against the allocation of blocks of numbers for particular types of material since we cannot foresee the extent to which various parts of the collection will grow. It is of no use to make a numerical principle for the arranging of the collection since one is always likely to find subsequent additions requiring the creation of multiple blocks and increasing the complexity of storage. To sum up, then, each language begins its tape numbering from 001 and continues to increase serial numbers indefinitely. Tapes can be shelved first by the language designation and thereafter in numerical order, or if desired alphabetically by the second letter of the prefix and numerically thereafter. It makes little difference in practice since we do not have an open access tape library and tapes are issued by a qualified assistant. A student merely needs to find the material he wants in the card index either by author, formal category or topic, to note the number and to ask the assistant for the appropriate tape.

From the card index system it is possible to compile registers with brief entries of what is available in the collection, and the form the registers take will depend on the needs of the departments interested e.g. literary items can be simply listed under each language in alphabetical order of author. Items of interest to sociologists can be listed under language and/or topic, and so on.

For administrative convenience we have adopted a colour coding system as a check that tapes and records do not become mislaid. Each language uses different coloured index cards and corresponding colour flashes on the spines of the tape boxes and on the spools themselves. In addition, the content of the tape is briefly noted inside the box and each spool is labelled with the tape number and its content; this information is also duplicated on the leader. Furthermore, as a safeguard against accidental erasure or other damage, original recordings are treated as archive material and are never normally used once they have been indexed and processed. These original recordings bear the suffix M

after the tape code reference to indicate an 'archive' master, while a second master tape, copied from the original, is designated by the suffix M_2. This second master is available for use by the teaching staff and is also the tape from which student copies will be made if required. Thus it is a relatively simple matter to make a new working copy (M_2) from the archive master should it become necessary. The economic disadvantage of having to keep a large and steadily increasing number of tapes lying idle is, we feel, justified and to some extent offset by the possibly greater cost which might arise in replacing lost material, if in fact it could be replaced at all.

In conclusion, it is not suggested that the tape classification system outlined above is ideal; merely that it appears to meet our particular needs at Essex, that it is relatively easy to operate and that it provides ready access both to the tapes themselves and to detailed notes of their content. We should like to express our thanks for their suggestions and help to our colleagues in the university library, especially to Mr. D. B. Butler, but for all inconsistencies and shortcomings we ourselves are entirely responsible.

Prof. J. B. Kay and A. Jameson
University of Essex
Colchester/Great Britain

APPENDIX

LIST OF TAPE REFERENCE CODES

(1) First Letter—Language
- F French
- G German
- P Portuguese
- R Russian
- S Spanish

(2) Second Letter—Formal Category
- (a) Author Index (Literature)
  - D Drama
  - P Prose
  - S Song
  - V Verse
- (b) Symbol Index (Non-literary Material)
  - Z Monologue: unprepared (novel speech)
  - Y Monologue: prepared but unscripted (exposition, etc.)

X Monologue: from notes (e.g. lecture)
W Monologue: fully prepared (address, etc.)
T Polylogue: unprepared (free discussion)
R Polylogue: one speaker prepared (interview, etc.)
Q Polylogue: both or all prepared (debate, etc.)

N Mixed formal content (e. g. news commentary/documentary)
C Course Material

(3) Number—Order of Access
Separate series starting with 001 for each language

(4) Abbreviated Subject in Tape Titles
(for non-literary recordings)

Biol.	Biology	Nat. Hist.	Natural History
Chem.	Chemistry	Perf. Arts	Performing Arts
Econ.	Economics	Plas. Arts	Plastic Arts
Hist. Occ.	Historic Occasions	Phil.	Philosophy
Hist.	History	Phys.	Physics
Law	Law	Pol.	Politics
Ling.	Linguistics	Soc.	Sociology
Med.	Medicine	Trav.	Travel
Milit.	War & Military		

Note: This list is being extended as the collection of recordings grows and is subject to revision.